Minería de Datos utilizando estrategias adaptativas Aplicaciones en optimización de procesos y modelización

Laura Lanzarini[1], Javier López[2], Waldo Hasperue[3], Leonardo Corbalán[4], María Delia Grossi[5], Juan Maulini[6], Augusto Villa Monte[7]

Instituto de Investigación en Informática LIDI (III-LIDI)[8]
Facultad de Informática. UNLP

CONTEXTO

Esta presentación corresponde al Subproyecto "Sistemas Inteligentes" perteneciente al Proyecto "Algoritmos Distribuidos y Paralelos. Aplicación a Sistemas Inteligentes y Tratamiento Masivo de Datos" del Instituto de Investigación en Informática LIDI.

RESUMEN

Esta línea de investigación se centra en el estudio y desarrollo de estrategias adaptativas que permitan resolver problemas de optimización de procesos y modelización de la información disponible.

Dentro de la Minería de Datos aplicada a la Optimización de Procesos, el énfasis está puesto en la búsqueda de reglas o patrones ocultos en históricos que puedan ayudar en la toma de decisiones o en la mejora de procesos productivos. En particular, se están investigando distintas metaheurísticas aplicables a la resolución del problema de ruteo de vehículos a fin de establecer un procedimiento de asignación de ambulancias y móviles a las prestaciones que realiza una empresa de emergencias médicas.

En lo referido a modelización de la información disponible, los temas centrales se encuentran relacionados con la investigación de nuevas estrategias adaptativas que generen agrupamientos a partir de grandes volúmenes de datos y además sean capaces de modificar su estructura ante algún cambio en los datos y/o nueva información que se obtenga, reflejando estos cambios en el conocimiento actual adquirido.

Las técnicas estudiadas y propuestas dentro de esta línea de investigación también han sido utilizadas en el ámbito de la docencia. En esta dirección, el énfasis está puesto en el análisis del material utilizado en las cátedras relacionadas con los miembros de este proyecto. En este trabajo se resume brevemente las tareas realizadas.

Palabras claves : Redes Neuronales, Algoritmos Evolutivos, Minería de Datos, Técnicas de Optimización.

1. INTRODUCCIÓN

En el Instituto de Investigación en Informática LIDI se está trabajando, desde hace varios años, en la resolución de problemas pertenecientes al área de Minería de Datos utilizando estrategias adaptativas. Se comenzó desarrollando diferentes mecanismos de aprendizaje y adaptación de redes neuronales competitivas que facilitaron la construcción del modelo de la información disponible principalmente a partir de reglas de asociación y clasificación [Has05, Has06]. Luego, con el objetivo de permitir que el modelo se adapte a los cambios del entorno, se desarrolló un método de obtención de reglas difusas [Has08]. También se propuso una solución para facilitar al usuario el uso de estas herramientas y se desarrolló una estrategia que define, a partir del modelo basado en reglas de clasificación, las acciones

[1] Profesor Titular. Facultad de Informática. UNLP
[2] Magister en Administración de Empresas.
[3] Becario de Postgrado Tipo I (CONICET) – Ayudante Diplomado - Facultad de Informática. UNLP
[4] Profesor Adjunto – Facultad de Informática – UNLP.
[5] Jefe de Trabajos Prácticos – Facultad de Ingeniería - UBA
[6] Becario III-LIDI. Ayudante Diplomado - Facultad de Informática. UNLP
[7] Becario III-LIDI.Ayudante Alumno - Facultad de Informática. UNLP
[8] Calle 50 y 115 1er Piso, (1900) La Plata, Argentina, TE/Fax +(54) (221) 422-7707. http://weblidi.info.unlp.edu.ar

a seguir para lograr el beneficio esperado [Has07].

Estas soluciones basadas en redes neuronales se complementan con las investigaciones realizadas sobre distintas metaheurísticas y en particular la inteligencia de cúmulos de partículas, a fin de poder controlar características especiales de la búsqueda con el objetivo de resolver problemas de predicción [Lan08a, Lop09].

A continuación se detallan brevemente los avances realizados últimamente.

1.1. Técnicas de Optimización aplicadas al ruteo de vehículos

Esta línea de trabajo está basada en la investigación y aplicación de metaheurísticas poblacionales para la resolución de problemas mono-objetivo y multi-objetivo. En particular, se selecciona la metaheuristica PSO (*Particle Swarm Optimization*), dado los buenos resultados alcanzados por la misma. Esta metaheuristica propuesta por Kennedy y Eberhart [Ken95] ha sido investigada en profundidad y se han propuesto variantes del algoritmo alcanzando resultados satisfactorios [Lan08a, Lop09].

Actualmente se busca aplicar PSO a una problemática del mundo real, de la rama de planificación de actividades y asignación de recursos. Específicamente, se utilizará este algoritmo, dentro de un proceso general de mejora de servicio en una empresa de emergencias médicas. Se propone un procedimiento de asignación de ambulancias y móviles a las prestaciones en forma automatizada.

Este proceso de vinculación entre las emergencias médicas a atender y los móviles y recursos asociados a cada prestación es la actividad central en este tipo de empresas.

El principal objetivo que se persigue consiste en lograr el máximo aprovechamiento de los recursos disponibles, sujetos a las restricciones de tiempo y económicas que caracterizan este tipo de problemas.

1.2. Modelización utilizando Redes Neuronales Competitivas

Esta línea de investigación está centrada en la definición de estrategias adaptativas que permitan extraer conocimiento de grandes bases de datos a partir de un modelo dinámico capaz de adaptarse a los cambios de la información, así como en el estudio de la optimización de la respuesta de los algoritmos a partir de su paralelización.

En especial se estudian métodos de clustering y clasificación de patrones para lograr asociar respuestas dinámicas con los datos de entrada obtenidos. Así, se espera conseguir métodos y técnicas de minería de datos que sean capaces de generar conocimiento útil, produciendo resultados que sean de provecho al usuario final.

Además, dado el gran volumen de información a procesar (que puede estar físicamente distribuida), resulta de interés investigar la arquitectura y paradigma de programación paralela utilizable de modo de minimizar el tiempo de cálculo del proceso adaptativo.

Los resultados de esta investigación pueden aplicarse en áreas tales como análisis de suelos, análisis genético, robótica, economía, medicina y comunicación de sistemas móviles. En estos casos es importante la obtención de un resultado óptimo, de modo de mejorar la calidad de las decisiones que se toman a partir del procesamiento. Desde el punto de vista informático estos problemas son un desafío interesante debido al volumen y distribución de los datos a analizar (incluso su complejidad) para obtener el conocimiento buscado.

Actualmente se está trabajando sobre una base de datos con información climática de todo el continente americano y sobre el área de distribución de 116 especies de triatominos. Los triatominos (más conocidos como vinchucas) son los principales vectores de la enfermedad del mal de Chagas. Y lo que se busca en esta base de datos son patrones de comportamiento de cómo las distintas especies fueron migrando a lo largo del

tiempo e intentar predecir cuales serían las próximas regiones que tendrán presencia de especies de vinchucas debido a los cambios climáticos que suceden hoy en día. Es decir, tratar de determinar las eventuales regiones (ciudades) que hoy por hoy no cuentan con presencia de vinchucas y que a futuro podrían llegar a tener. Con estos resultados, los biólogos, podrían determinar donde podrían migrar las distintas especies y eventualmente plantear como evitar dicha migración.

También se están utilizando redes neuronales competitivas para resolver el problema de la búsqueda de recursos en sistemas distribuidos Peer-to-Peer (P2P). En esta dirección, y como resultado preliminar de esta línea de investigación se ha presentado una estrategia basada en la propagación selectiva de las solicitudes de búsqueda únicamente a los vecinos más adecuados de los nodos [cor09a].

Actualmente se están ensayando estrategias alternativas al aprendizaje tradicional de las redes neuronales que intervienen en las decisiones de búsqueda en cada nodo de la red P2P. El objetivo es acelerar los tiempos de entrenamiento y mantener actualizado el conocimiento adquirido por cada nodo. Para ello se está trabajando en una solución multiagente basada en la metaheurística Colonias de Hormigas (ACO) encargada de llevar adelante un apropiado intercambio de información entre los nodos de la red.

El diseño de métodos de exploración incremental, implementados como sucesiones de búsquedas limitadas dirigidas a distintos nodos de la red cuyo alcance de propagación va ampliándose gradualmente, también constituye un tema de estudio en la presente línea de trabajo [cor09b].

1.3. Minería de Datos en Educación

La aplicación de técnicas de Minería de Datos en el ámbito educativo ha permitido caracterizar a los distintos actores que intervienen en los procesos de enseñanza-aprendizaje [ROM05]. En particular, durante 2009 y partir de tareas realizadas en forma conjunta entre el Instituto de Investigación en Informática LIDI (UNLP) y el Centro de Investigación en Procesos Básicos, Metodologías y Educación – UNMDP) se han realizado contribuciones referidas al Modelo de Estudiante [Lan09, Aor09].

Actualmente se continúa estudiando en la aplicación de técnicas de Minería de Datos con el objetivo de evaluar la pertinencia y calidad del material desarrollado para un curso dado [Lan08b, Gro08]. Uno de los problemas centrales encontrados hasta el momento es la recolección centralizada de la información referida a la participación e interacción del alumno con dicho material. El conocimiento por parte de los docentes de los pasos recorridos por los alumnos durante el proceso de enseñanza-aprendizaje es de suma importancia para la toma de decisiones en la metodología a utilizar [GON04].

Por tal motivo, se ha propuesto el desarrollo y la utilización de una herramienta de software que monitoree este proceso en una de las etapas más importante: la realización de los trabajos prácticos. Con esto se espera poder contar con información permanentemente actualizada del desempeño de los alumnos y a la vez incorporar una metodología de trabajo que favorezca tanto a alumnos como docentes.

Los datos recolectados serán utilizados para obtener reglas de asociación que permitan mejorar el material y/o los procesos de aprendizaje.

2. TEMAS DE INVESTIGACIÓN Y DESARROLLO

- Desarrollo e implementación, a partir de los métodos existentes, de estrategias adaptativas capaces de construir y mantener modelos adecuados en entornos de información dinámicos.
- Análisis de los distintos tipos de Redes Neuronales competitivas dinámicas. Estudio de las estrategias existentes que permiten determinar, durante la adaptación, el tamaño de la arquitectura y

forma de conexión de los elementos que componen la red neuronal.

- Estudio y aplicación de diferentes de métricas que permitan analizar la preservación de la topología de los datos tanto en el espacio de los patrones de entrada como en el espacio de salida de la red.
- Estudio, desarrollo y aplicación de variantes de PSO que faciliten su aplicación a la resolución del problema de ruteo de vehículos y en especial a la asignación de ambulancias de una empresa de emergencias médicas.
- Educación basada en la WEB. Relación entre procesos cognitivos e informáticos.
- Estudio e investigación de Sistemas Hipermedia Adaptativos basados en WEB existentes.

3. RESULTADOS OBTENIDOS/ ESPERADOS.

- Desarrollo e implementación de una estrategia para la obtención de reglas de clasificación a partir de los hipercubos definidos mediante una estrategia de clustering.
- Desarrollo e implementación de una estrategia basada en cúmulos de partículas (PSO) que permita resolver el problema de asignación de Asignación de recursos en forma optimizada, reduciendo tiempos de llegada a las diferentes prestaciones, aumentando la calidad de servicio y maximizando el rendimiento de los recursos disponibles. Esto permitirá:
  - Incrementar la información disponible para los responsables de asignación de móviles y seguimiento de prestaciones.
  - Estandarizar el proceso de alocación de móviles a prestaciones, considerando el estado de situación completa.
  - Es una herramienta de operación, que complementa la actividad de los responsables y permite liberar tiempo de estos para actividades de mayor valor agregado.
  - Mejorar la asignación manual existente.
- Diseño e implementación del material didáctico del curso cuya pertinencia y calidad se desea evaluar.
- Aplicación de técnicas de Minería de Datos a la información extraída de sistemas educativos para la posterior evaluación de los mismos.
- Mejoramiento de los sistemas de enseñanza utilizados

4. FORMACIÓN DE RECURSOS HUMANOS

Dentro de los temas involucrados en esta línea de investigación se están desarrollando actualmente 2 tesis de doctorado, 1 de maestría y al menos 2 tesinas de grado de Licenciatura. También participan en el desarrollo de las tareas becarios y pasantes del III-LIDI.

5. REFERENCIAS

[Aor09] “Lógica Difusa aplicada al Modelo del Estudiante de un Sistema Tutorial Inteligente”. Arona, Huapaya, Lanzarini, Lanzarini, Lizarralde. IV Congreso de Tecnología en Educación y Educación en Tecnología. TE&ET’09. Julio 2009. La Plata. Bs.As

[Cor09a] “Resources NeuroSearch in Peer-to-Peer Networks”. Corbalán, Lanzarini, De Giusti. 31st International Conference on Information Technology Interfaces. University of Zagreb, University Computing Centre. Cavtat/ Dubrovnik, Croatia. June 2009. pp 597 – 602.

[Cor09b] “Búsqueda Neuronal de Recursos con Exploración Incremental en Redes Peer-to-Peer”. Corbalán, Lanzarini, De Giusti. Jornadas Chilenas de Computación. JCC 2009. Santiago de Chile. Noviembre 2009. XIII Workshop on Distributed Systems and Parallelism (WSDP). pp 11-20.

[GON04] Carina Soledad González. Sistemas inteligentes en la educación: una revisión de las líneas de investigación y aplicación actuales. Revista Electrónica de Investigación y Evaluación Educativa RELIEVE, 10(1):3-22, 2004.

[Gro08] "Reglas de Predicción aplicables al Diseño de un Curso de Computación". Grossi, Lanzarini. III Congreso de Tecnología en Educación y Educación en Tecnología. TE&ET'08. Mayo 2008. Bahía Blanca.

[Has05] Hasperué, Lanzarini. Dynamic Self-Organizing Maps. A new strategy to enhance topology preservation. *XXXI Conferencia Latinoamericana de Informática*. CLEI 2005.

[Has06] Hasperué, Lanzarini. Classification Rules obtained from Dynamic Self-organizing Maps. *VII Workshop de Agentes y Sistemas Inteligentes*. CACIC 2006. San Luis. Argentina. Octubre de 2006.

[Has07] Hasperué, Lanzarini. Extracting Actions from Classification Rules. *Workshop de Inteligencia Artificial. Jornadas Chilenas de Computación 2007*. Iquique, Chile. Noviembre de 2007.

[Has08] Hasperué, W., G. Osella Massa & L. Lanzarini. Obtaining a Fuzzy Classification Rule System from a non-supervised Clustering. *30th International Conference of Information Technology Interfaces (ITI)*. Cavtat, Croacia. June 2008.

[Ken95] Kenedy, Eberhart. Particle Swarm Optimization. *IEEE International Conference on Neural Networks*. Vol IV, pp.1942-1948. Australia 1995.

[Lan00] Lanzarini. Reconocimiento de Patrones en Imágenes Médicas utilizando Redes Neuronales. *Journal of Computer Science and Technology*. Vol.4 . Dic 2000.

[Lan04] Lanzarini, Yanivello. Reconocimiento de Comandos Gestuales utilizando GesRN. *V Workshop de Agentes y Sistemas Inteligentes*. CACIC 2004. Bs.As. Argentina. 2004. ISBN 987-9495-58-6

[Lan08b] "Estrategias Inteligentes aplicables a un Sistema Educativo". Lanzarini, Denazis, Grossi. X Workshop de Investigadores en Ciencias de la Computación (WICC 2008), Area Tecnología Informática Aplicada en Educación. Mayo de 2008. La Pampa

[Lan08a] Lanzarini, Leza, De Giusti. Particle Swarm Optimization with Variable Population Size. *Lecture Notes in Computer Science*. Vol 5097/2008. Artificial Intelligence and Soft Computing. Pags. 438-449. Junio de 2008. ISBN 987-3-540-69572-1.

[Lan09] "Diagnóstico Adaptativo del Estudiante en Sistemas Tutoriales Inteligentes".Lanzarini, Huapaya. XI Workshop de Investigadores en Ciencias de la Computación (WICC 2009), Area Tecnología Informática Aplicada en Educación. Mayo de 2009. San Juan

[Lop09] "Particle Swarm Optimization with Oscillation Control". Lopez, Lanzarini, De Giusti. Genetic and Evolutionary Computation Conference. ACM GEECCO Proceeding. Montréal, Canada. July 2009.

[ROM05] Cristóbal Romero Morales, Sebastián Ventura Soto, Cesar Hervás Martínez. Estado actual de la aplicación de la minería de datos a los sistemas de enseñanza basada en web. Actas del III Taller Nacional de Minería de Datos y Aprendizaje, TAMIDA2005, pp.49-56. ISBN: 84-9732-449-8

[Ken95] Kennedy J., Eberhart R. Particle Swarm Optimization, in Proceedings of the IEEE International Joint Conference on Neural Networks, pp 1942-1948, IEEE Press, 1995

[Shi99] Shi Y., Eberhart R. An empirical study of particle swarm optimization. *IEEE Congress Evolutionary Computation*. pp.1945-1949. Washington DC, 1999.